STATISTIQUES DESCRIPTIVES

CH1 : AJUSTEMENT ET CORRELATION LINEAIRE

SERIE STATISTIQUE A DEUX VARIABLES

- Une série statistique à deux caractères quantitatifs, xi et yi, est **une série double** dont les valeurs sont données par les couples (xi; yi) .
- L'ensemble de ces points forme un nuage de points. Ce nuage peut avoir une forme allongée, curviligne ou très dispersée.
- **Point moyen du nuage**

On appelle **point moyen G(x; y)** le point dont les coordonnées sont les moyennes des valeurs xi et yi de la série.

$$x_G = \frac{\sum xi}{N}, \quad y_G = \frac{\sum yi}{N}$$

N est l'effectif total des observations pour les deux séries statistiques

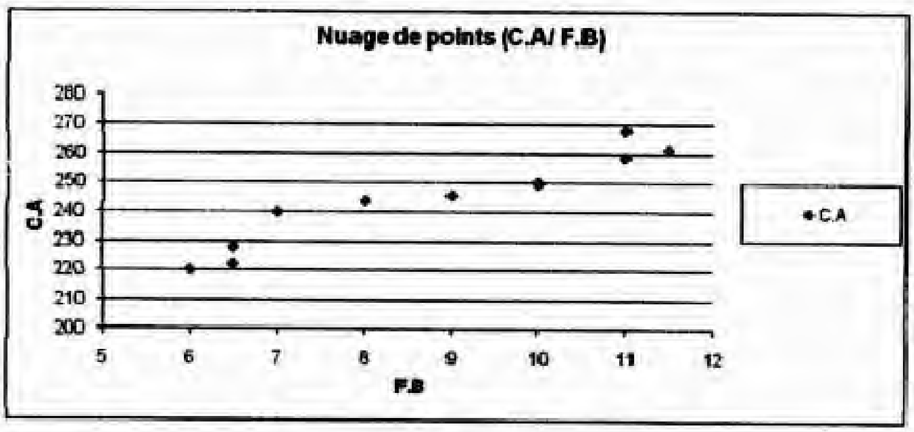
Séries statistiques bivariées

- **Exemple**

Un responsable de ventes de magasin analyse l'évolution de son chiffre d'affaires sur la dernière période. Il relève pour cela le montant des frais de publicité engagés sur la même période. Il dresse le tableau suivant (les montants sont exprimés en centaines d'euros)

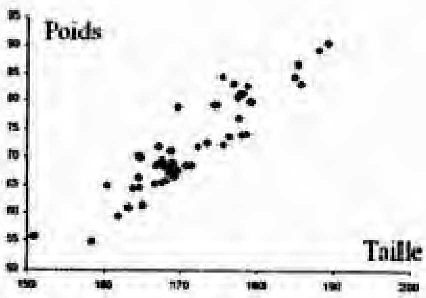
Frais de publicité X_i	10	6	6,5	11,5	11	8	7	6,5	11	9
Chiffre d'affaires y_i	250	220	228	262	268	244	240	222	259	246

EXEMPLE 1



EXEMPLE 2

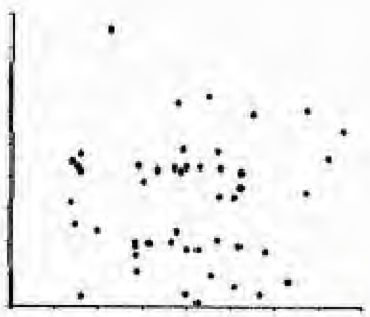
Nom	Taille x_i (cm)	Poids y_i (kg)
ALI	175	73
AHMED	168	56
......	......	......
SALMA	185	87



La connaissance de la taille x apporte une certaine information sur le poids y

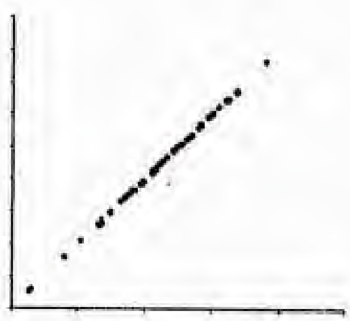
Il existe une <u>relation de dépendance</u> entre x et y

Cas possibles:



La connaissance de x n'apporte aucune certaine information sur y

x et y sont <u>indépendantes</u>



La connaissance de x permet de connaitre exactement la valeur de y

Il existe une <u>relation fonctionnelle</u> entre x et y

$A\binom{}{}$ $B\binom{}{}$ $y = 18x + 100$

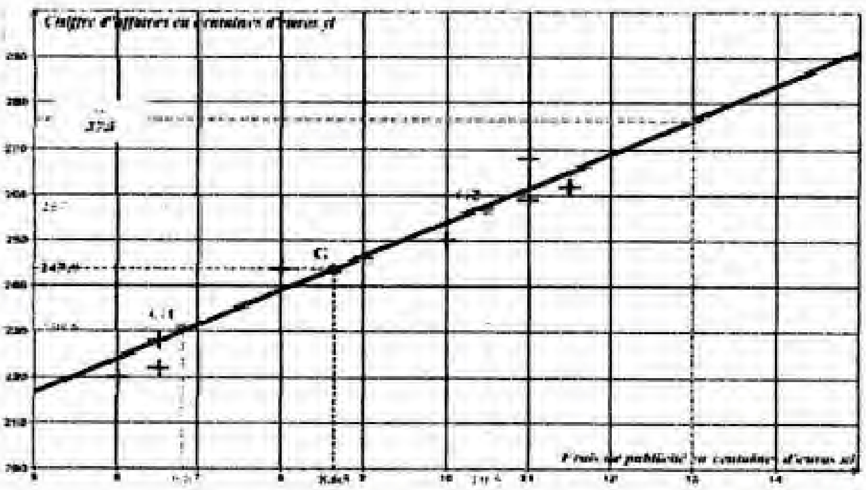
$y = ax + b$

Ajustement linéaire (méthode de Mayer)

- Dans le cas d'un nuage de points de forme allongée, et afin de faciliter l'étude de la série, il est possible de représenter ce nuage par une droite appelée droite d'ajustement linéaire.
- Pour tracer cette droite, on utilise la méthode de Mayer.
- Le nuage est partagé suivant les valeurs croissantes de xi en deux nuages d'égale importance :
    - **on calcule les coordonnées des points moyens G1 et G2 de ces deux nuages;**
    - **on détermine l'équation de la droite (Gl G2).**
- Cette droite est appelée **droite de Mayer.**
- Elle passe par le point moyen G.

Revenons à l'exemple précédent.

On peut construire une droite d'ajustement linéaire par la méthode de Mayer



Covariance de deux variables statistiques

DEFINITION DE LA COVARIANCE:

$$\mathrm{Cov}(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

On simplifie cette formule pour retrouver la suivante:

$$Cov\,(x,y) = \overline{xy} - \bar{x}\,\bar{y}$$

Propriétés :

$\mathrm{Cov}(x,y) > 0 \Leftrightarrow$ x et y varient dans le même sens

$\mathrm{Cov}(x,y) < 0 \Leftrightarrow$ x et y varient en sens contraire

$\mathrm{Cov}(x,y) = \mathrm{Cov}(y,x)$

$\mathrm{Cov}(x,x) = V(x)$

$\mathrm{Cov}(ax+by\,,z) = a\,\mathrm{Cov}(x,z) + b\,\mathrm{Cov}(y,z)$

Exemple:

- Revenons à l'exemple précédent:

Frais de publicité Xi	10	6	6,5	11,5	11	8	7	6,5	11	9	$\bar{X}$ 8,65
Chiffre d'affaires yi	250	220	228	262	268	244	240	222	259	246	$\bar{y}$ 243,9
XiYi	2500	1320	1428	3013	2948	1952	1680	1443	2849	2214	$\overline{XY}$ 2140,1

COV(X,Y)=30,36

X et Y sont deux variables qui varient dans le même sens.

MESURE DE L'INFLUENCE MUTUELLE DE DEUX VARIABLES STATISTIQUES (Corrélation linéaire)

Corrélation linéaire: $\rho = \dfrac{\text{cov}(x,y)}{\sigma(x)\,\sigma(y)}$

__Propriétés :__

$$-1 \leq \rho \leq 1$$

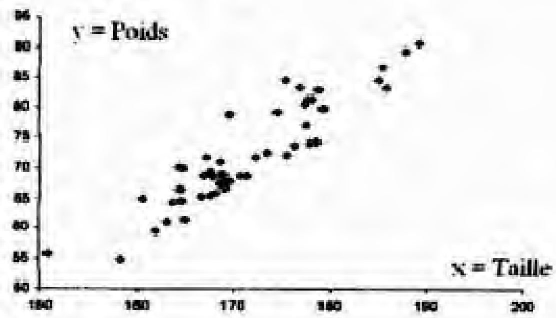
$$y = a\,x + b \Leftrightarrow \begin{cases} \rho = 1 & \text{si } a > 0 \\ \rho = -1 & \text{si } a < 0 \end{cases}$$

$|\rho| = 1 \Leftrightarrow$ Il existe une relation fonctionnelle entre x et y

$\rho = 0 \Leftarrow$ x et y sont independants

$0 < |\rho| < 1 \Leftrightarrow$ Il existe une dépendance linéaire d'autant plus forte que $|\rho|$ est grand

AJUSTEMENT LINEAIRE

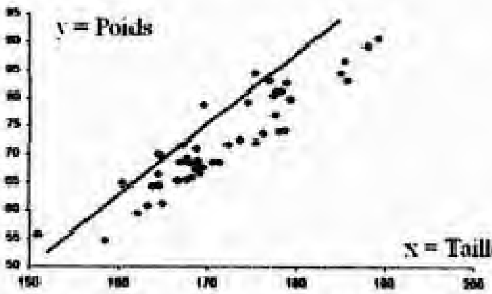


Est-il possible de trouver une fonction numérique f telle que $y = f(x)$?

Si une telle fonction existe, on dit que f est un modèle du phénomène étudié.

x est la variable explicative.
y est la variable expliquée.

AJUSTEMENT LINEAIRE



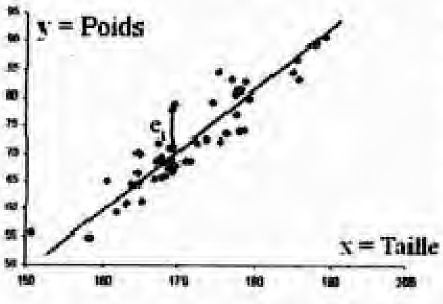
On désire trouver la droite qui passe « au mieux » à l'intérieur du nuage de points

AJUSTEMENT LINEAIRE

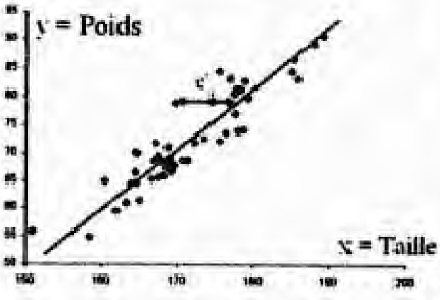
« au mieux »

Minimiser $S = \sum_{i=1}^{n} e_i^2$

Minimiser $S' = \sum_{i=1}^{n} e_i'^2$



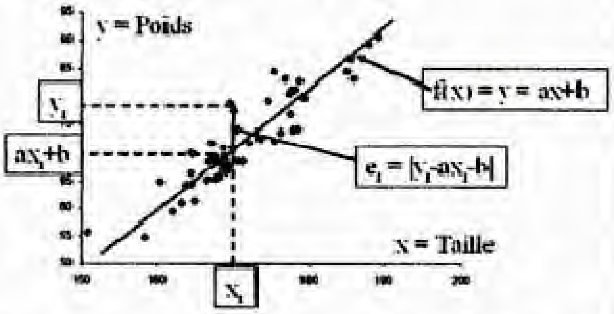
Droite de régression de y en x



Droite de régression de x en y

AJUSTEMENT LINEAIRE
REGRESSION LINEAIRE DE Y EN X

Droite de régression linéaire de y en x
$y = f(x) = ax + b$



La droite de régression linéaire de y en x, notée $D_{y,x}$, minimise $S = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - ax_i - b)^2$

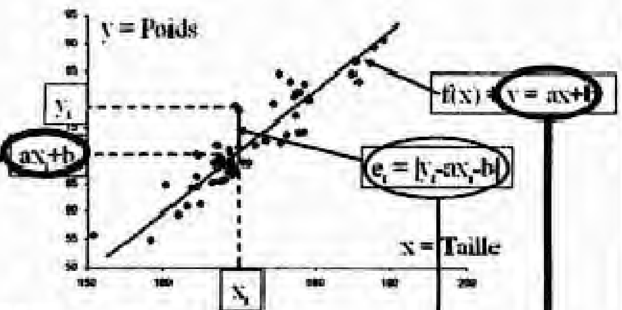
$$a = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{Cov(x,y)}{V(x)}$$

$$\boxed{b = \bar{y} - a\bar{x}}$$

$D_{y,x}$ passe par le point moyen $(\bar{x}, \bar{y})$

REGRESSION LINEAIRE DE Y EN X

Droite de régression linéaire de y en x
$y = f(x) = ax + b$



$y = a x + b$ définit un modèle affine

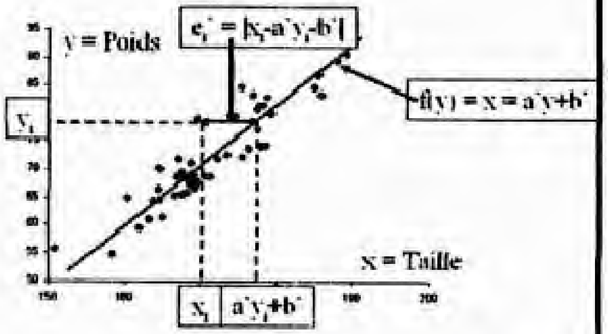
$\hat{y}_i = a x_i + b$ = valeur de y_i prévue par le modèle

$r_i = y_i - \hat{y}_i$ = résidu de la ième observation

$e_i = |r_i| = |y_i - a x_i - b|$ = erreur due au modèle

REGRESSION LINEAIRE DE X EN Y

Droite de régression *linéaire de x en y*
$x = f(y) = a'y + b'$



La droite de régression linéaire de x en y, notée $D_{x,y}$, minimise $S' = \sum_{i=1}^{n} e_i'^2 = \sum_{i=1}^{n} (x_i - a'y_i - b')^2$

$$a' = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(y_i-\bar{y})^2} = \frac{Cov(x,y)}{V(y)}$$

$$\boxed{b' = \bar{x} - a'\bar{y}}$$

$D_{x,y}$ passe par le point moyen $(\bar{x}, \bar{y})$

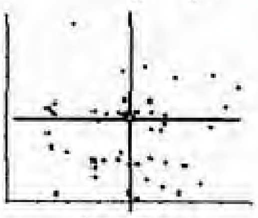
LIENS ENTRE CORRELATION ET DROITES DE REGRESSION

$D_{y,x}$: $y = ax + b$ $\qquad a = \frac{Cov(x,y)}{V(x)}$ $\qquad b = \bar{y} - a\bar{x}$

$D_{x,y}$: $x = a'y + b'$ $\qquad a' = \frac{Cov(x,y)}{V(y)}$ $\qquad b' = \bar{x} - a'\bar{y}$

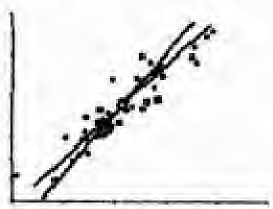
$\Leftrightarrow y = \frac{1}{a'}x - \frac{b'}{a'}$

$\rho^2 = a\,a'$ $\qquad \boxed{\rho = a\frac{\sigma(x)}{\sigma(y)} = a'\frac{\sigma(y)}{\sigma(x)}}$



$\rho^2 = a\,a' = 0$

Indépendance linéaire



$0 < \rho^2 = a\,a' < 1$

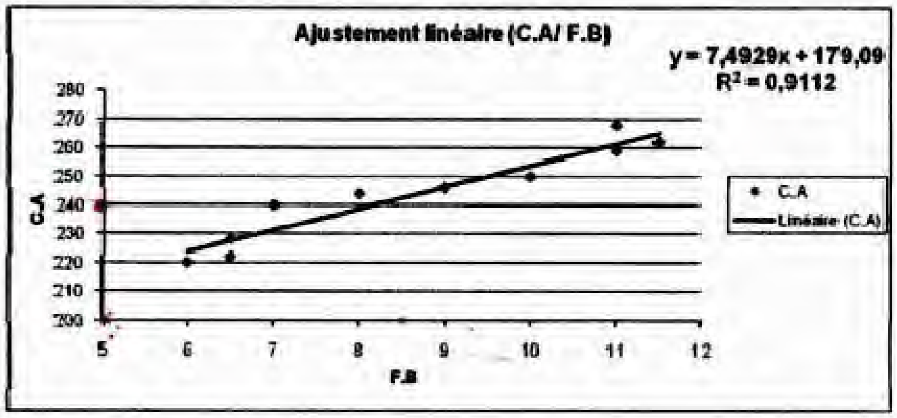
Le degré de dépendance linéaire se mesure à la proximité des droites de régression

$\rho^2 = a\,a' = 1$

Liaison fonctionnelle linéaire

[redacted]

Cherchons la droite d'ajustement linéaire par la méthode des moindres carrées pour l'exemple précédent:



[redacted]

Exemple:

Une distribution statistique se présente de la façon suivante :

Xi	2	3	6	7	9	10	11	12
Yi	21	22	15	14	10	8	4	2

1) Représentez graphiquement le nuage des points.
2) *Ajustez par la méthode de Mayer.*
3) Deux statisticiens différents ont abouti aux droites d'ajustements suivantes :

$$D_1 : y = -2x + 27{,}2.$$

$$D_2 : y = -1{,}9x + 26$$

- [redacted]
  - En se fondant sur le principe des moindres carrées, apprécier lequel des deux statisticiens a réalisé le meilleur ajustement. On calculera par les deux fonctions obtenues le total (yi-yi ajusté)² et on comparera les résultats obtenus.
  - Effectuer ensuite l'ajustement par la méthode des *moindres carrées*.
  - Vérifier à partir de l'équation de la droite des moindres carrées que le total (yi-yi ajusté) est nul.
  - Vérifier à partir de cette même équation de la droite des moindres carrées que le total (yi-yi ajusté)² est plus faible que chacun des deux totaux calculés.